Теоретико-групповой подход в комбинаторной теории оценок обобщающей способности

Фрей Александр Ильич

Московский физико-технический институт
(Государственный университет)
Кафедра «Интеллектуальные Системы» (ВЦ РАН)

Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович

25 апреля 2013

Теоретико-групповой подход к проблеме переобучения

2010 - 2012

- ✓ предложен рандомизированный метод обучения (РМЭР);
- ✓ разработан теоретико-групповой метод оценки вероятности переобучения РМЭР для модельных семейств;
- ✓ получен метод порождающих и запрещающих множеств для РМЭР;
- ✓ предложена рандомизация на множестве алгоритмов как модель отсутствия связности;
- ✓ предложена рандомизация на целевых метках и доказана теорема о декомпозиции профиля расслоения-связности;

Публикации и выступления 2010-2012

- 2009, ММРО, Фрей А. И., Точные оценки вероятности переобучения для симметричных семейств алгоритмов
- **2010, PRIA, Фрей А. И., Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения**
- 2010, ИОИ, Фрей А. И., Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов
- 2011, ММРО, Фрей А. И., Метод порождающих и запрещающих множеств для рандомизированного метода минимизации эмпирического риска
- 2012, EURO Conference on Operational Research, Frey A., Geometrical properties of connected search spaces for binary classification problem

Теоретико-групповой подход к проблеме переобучения

2012-2013

- ✓ получены новые экспериментальные результаты о переобучения логических закономерностей;
- ✓ предложена новая общая оценка вероятности переобучения, учитывающая сходство алгоритмов;
- ✓ проведено сравнение комбинаторных оценках с последними оценками PAC-bayes подхода (на примере логистической регрессии)

Публикации и выступления 2012-2013

- 2012, ИОИ, Фрей А. И., Ивахненко А. А, Решетняк И. М., Применение комбинаторных оценок вероятности переобучения в простом голосовании конъюнкций
- (на стадии подготовки) Фрей А. И, Толстихин И.О., Учет сходства алгоритмов в комбинаторной теории оценок обобщающей способности
- (на стадии подготовки) Соколов Е., Фрей А. И., Применение комбинаторных оценок вероятности переобучения при настройке логистической регрессии.
- (на стадии подготовки) Фрей А. И, Решетняк И. М., Учет верхней связности в комбинаторных оценках вероятности переобучения.
- (на стадии подготовки) K.V.Vorontsov, E.Sokolov, N. Zhivotovskiy, A.Frei., Combinatorial generalization bounds. (in Springer?)

Комбинаторный подход к проблеме переобучения

- Строки таблицы $\{x_1 \ldots x_\ell, x_{\ell+1}, x_L\}$ — объекты полной выборки
- Столбцы $\{a_1 \ldots a_D\}$ — векторы ошибок алгоритмов

	a_1	a_2	$\ldots$	a_d	$\ldots$	a_D
x_1	0	1	$\ldots$	**0**	$\ldots$	1
$\ldots$	1	1	$\ldots$	**0**	$\ldots$	0
x_ℓ	0	0	$\ldots$	**0**	$\ldots$	0
$x_{\ell+1}$	1	1	$\ldots$	**1**	$\ldots$	1
$\ldots$	1	0	$\ldots$	**1**	$\ldots$	0
x_L	0	0	$\ldots$	**1**	$\ldots$	0

- Метод обучения — минимизация эмпирического риска
- Цель: получить точные, вычислительно-эффективные оценки вероятности переобучения.

Определения

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное множество объектов,
A — множество алгоритмов,
$\mathbb{X} = X \sqcup \bar{X}$ — разбиение $\mathbb{X}$ на обучение и контроль,
$\nu(a, X) = \frac{1}{\ell} \sum\limits_{x_i \in X} I(a(x_i), y_i)$ — доля ошибок a на X,
$\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ — *переобученность* a *на* X,
$\mu X = \operatorname{argmin}\limits_{a \in A} \nu(a, X)$ — минимизация эмпирического риска,
Вероятность переобучения (ВП):

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} [\delta(\mu X, X) \geq \varepsilon].$$

ВП рандомизированной минимизации эмпирического риска:

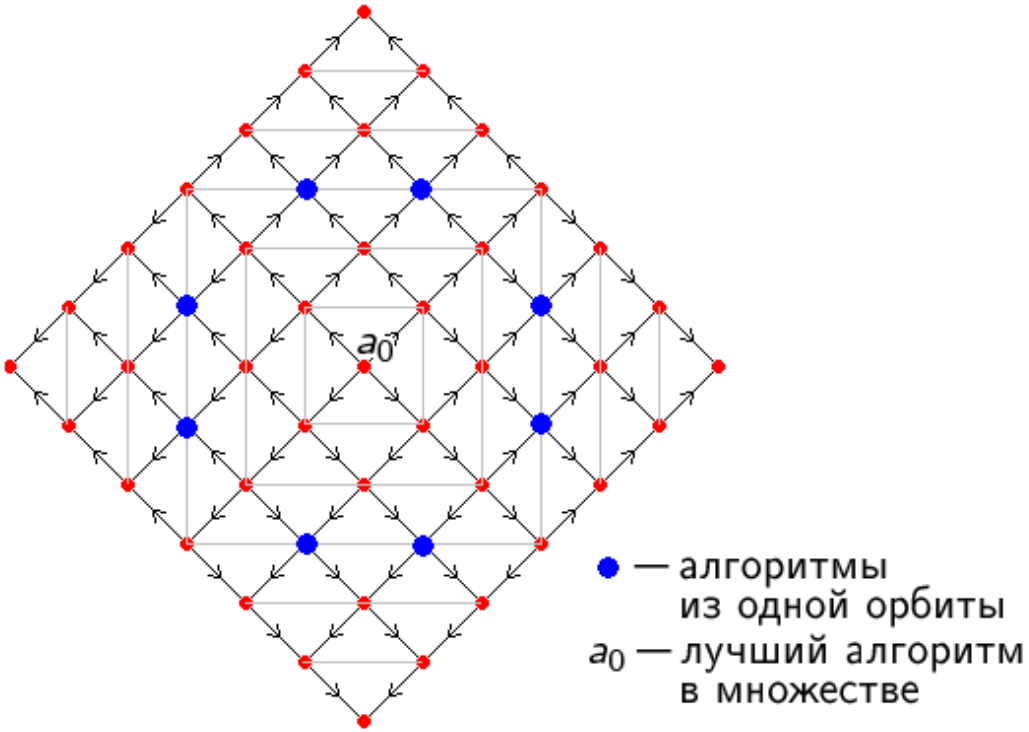
$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} \sum_{a \in A(X)} \frac{[\delta(a, X) \geq \varepsilon]}{|A(X)|}, \text{ где } A(X) = \operatorname{Argmin}_{a \in A} \nu(a, X).$$

Методы вывода формул для вероятности переобучения

1. Метод производящих и запрещающих объектов
   - Монотонная цепочка и сетка
2. Блочная оценка
   - Пара алгоритмов
3. Рекуррентное вычисление вероятности переобучения по заданной матрице ошибок
   - Теоретический инструмент для доказательства универсальных оценок
4. Гипотеза t-слоев и метод β-многочленов
   - Унимодальные цепочки и монотонные сети (точно)
   - Унимодальные сети (приближенно)
5. Метод разбиения множества алгоритмов на орбиты
   - [А. Фрей], Пучок монотонных цепочек
   - [А. Фрей], Полный слой, полный куб алгоритмов
   - [И. Толстихин], Шар алгоритмов и центральный слой
   - [А. Фрей], Монотонные и унимодальные сети (точно)

Группа симметрий множества алгоритмов

Граф смежности двумерной унимодальной сетки:



- S_L — группа всех перестановок объектов выборки,
- S_L действует множестве всех алгоритмов $2^{\mathbb{A}}$,
- $\mathrm{Sym}(A) = \{\pi \in S_L : \pi A = A\} \subset S_L$.
- Орбита алгоритма a это $\{\pi a : \pi \in \mathrm{Sym}(A)\} \subset A$

Равный вклад алгоритмов одной орбиты

- Вероятность переобучения — сумма вкладов алгоритмов:

$$Q_\varepsilon = \sum_{a \in A} Q_\varepsilon(a), \text{ где}$$

$$Q_\varepsilon(a) = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} \frac{[\delta(a, X) \geq \varepsilon]}{|A(X)|};$$

- Алгоритмы одной орбиты дают равный вклад:

$$Q_\mu(\varepsilon, a, A) = Q_\mu(\varepsilon, \pi a, A), \text{ где } \pi \in \mathrm{Sym}(A)$$

- Обозначим $\Omega(A)$ — множество орбит $\mathrm{Sym}(A)$ на A;
- Вероятность переобучения с учетом структуры множества алгоритмов:

$$Q_\varepsilon = \sum_{\omega \in \Omega(A)} |\omega| Q_\varepsilon(a_\omega). \qquad (1)$$

РМЭР: метод порождающих и запрещающих множеств

ММРО-2011, Фрей А.
"Метод порождающих и запрещающих множеств для рандомизированного метода минимизации эмпирического риска".

Theorem

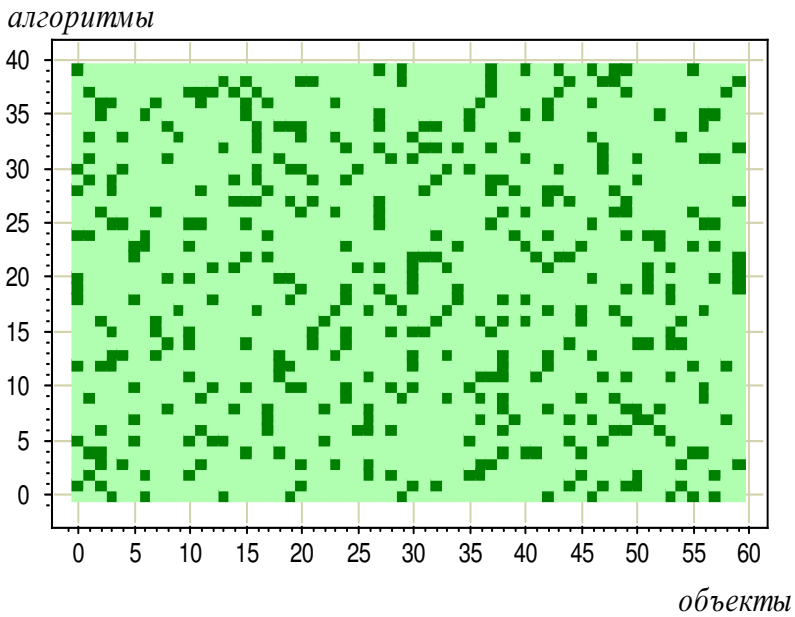
Введем множество $\aleph = \{A(X) : X \in [\mathbb{X}]^\ell\}$.
Пусть для каждого $\alpha \in \aleph$ *существуют порождающее и запрещающее множества* X *и* X', *такие что:*

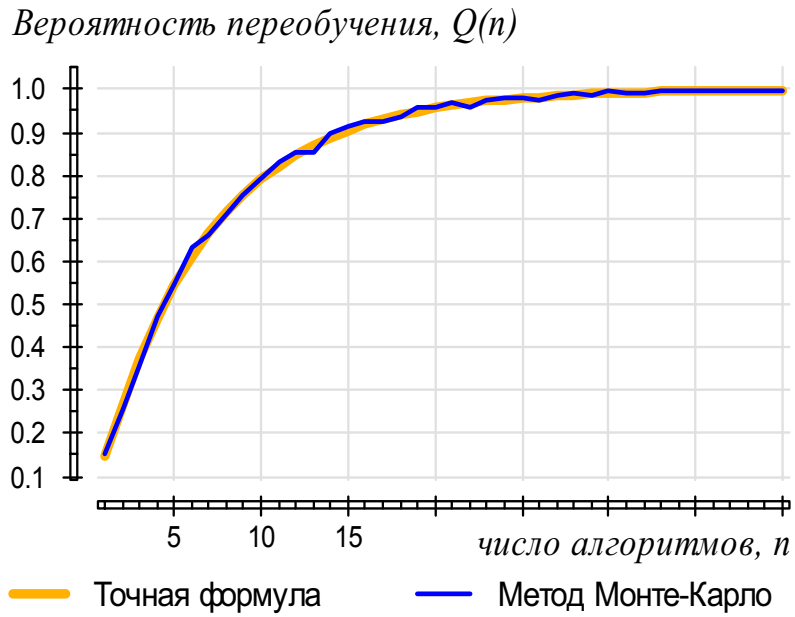
$$[A(X){=}\alpha] = [X_\alpha \subseteq X][X'_\alpha \subseteq \bar{X}], \quad \forall X \in [\mathbb{X}]^\ell.$$

Тогда ВП записывается в виде:

$$Q_\varepsilon(A) = \sum_{a\in A}\sum_{\alpha\in\aleph} \frac{[a\in\alpha]}{|\alpha|}\frac{C_{L_\alpha}^{\ell_\alpha}}{C_L^\ell} H_{L_\alpha}^{\ell_\alpha, m_\alpha^a}(s_\alpha^a(\varepsilon)).$$

Модель семейства без расслоения и без связности





- Пусть A_m^n — множество из n алгоритмов, допускающих по m ошибок. Векторы ошибок независимы.

Теорема (Вероятность переобучения для A_m^n)

Пусть μ — рандомизированный МЭР. Тогда

$$\bar{Q}_\varepsilon(A_m^n) = 1 - \left(1 - Q_\varepsilon(A_m^1)\right)^n$$

Декомпозиция профиля расслоения-связности

EURO-2012, 25th Conference on Operational Research. "Geometrical properties of connected search spaces for binary classification problem".

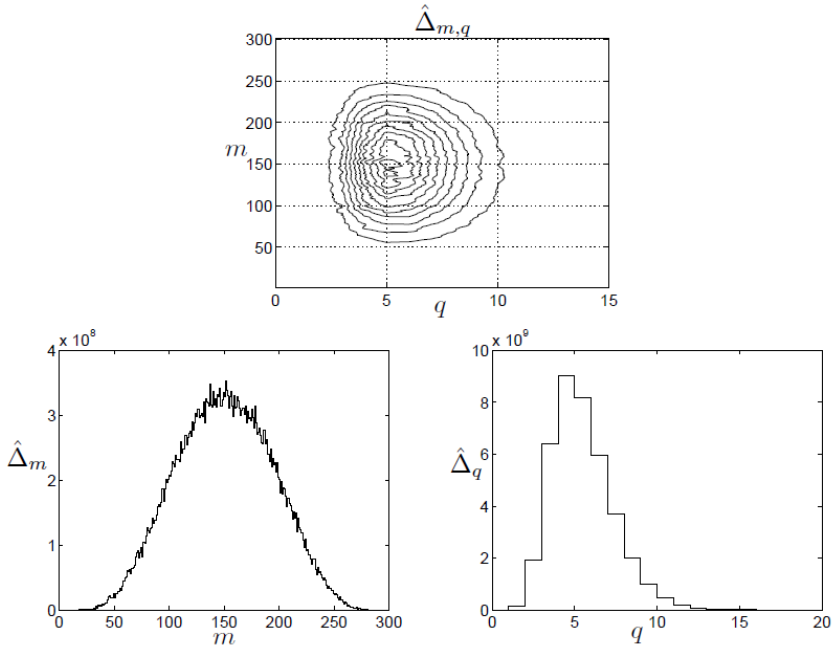
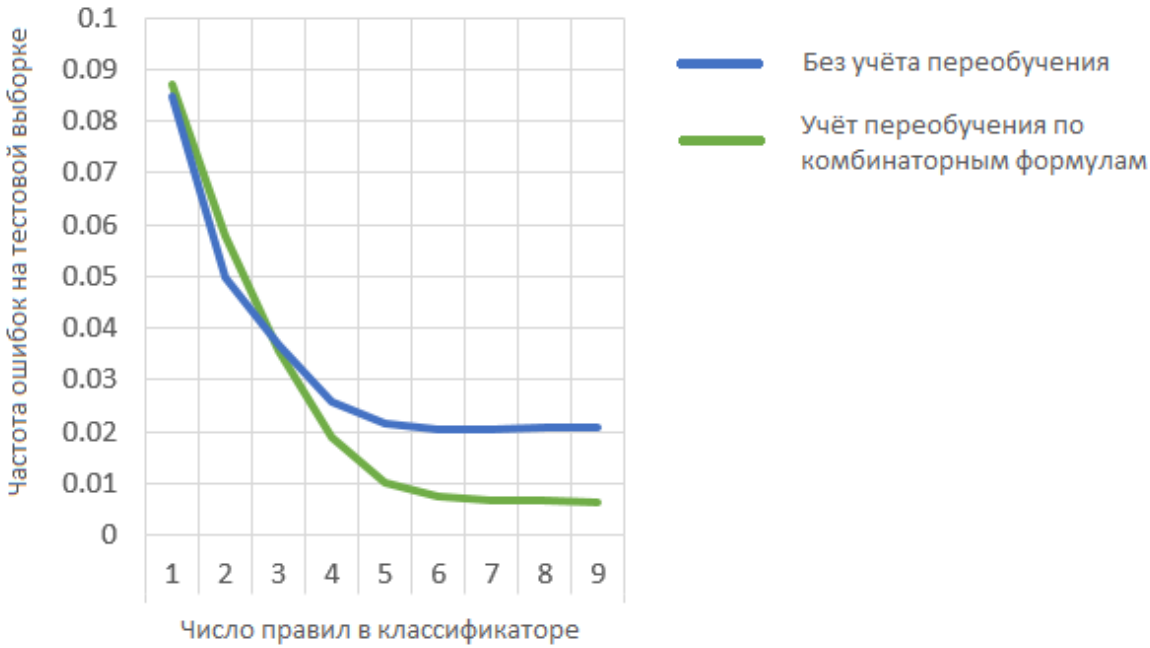


Рис. : Профиль расслоения-связности для семейства линейных классификаторов в $\mathbb{R}^p$. $p = 5$, $L = 300$, $|R| = 2 \cdot 10^5$.

Численный эксперимент

ИОИ-2012, А.Ивахненко, А.Фрей.
"Применение комбинаторных оценок вероятности переобучения в голосовании пороговых конъюнкций."



Зависимость частоты ошибок на тестовой выборке от числа правил в классификаторе. Задача Echo Cardiogram.

Верхняя оценка вероятности переобучения

Theorem (Воронцов, Решетняк, Ивахненко, 2010)

Для любого монотонного метода μ, любых $\mathbb{X}$, A и $\varepsilon \in (0,1)$

$$Q_\varepsilon(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} H_{L-u-q}^{m-q,\ell-u}\left(\frac{\ell}{L}(m-\varepsilon k)\right),$$

где u — верхняя связность алгоритма a,
q — неполноценность алгоритма a,
$m = m(a, \mathbb{X})$ — число ошибок алгоритма a.

Сходство алгоритмов - пример 1

Множество $A = (a_1, a_2)$ состоит из двух алгоритмов,
оба алгоритма допускают по m ошибок на полной выборке,
хэммингово расстояние $\rho(a_1, a_2) = d$ заранее фиксированно.

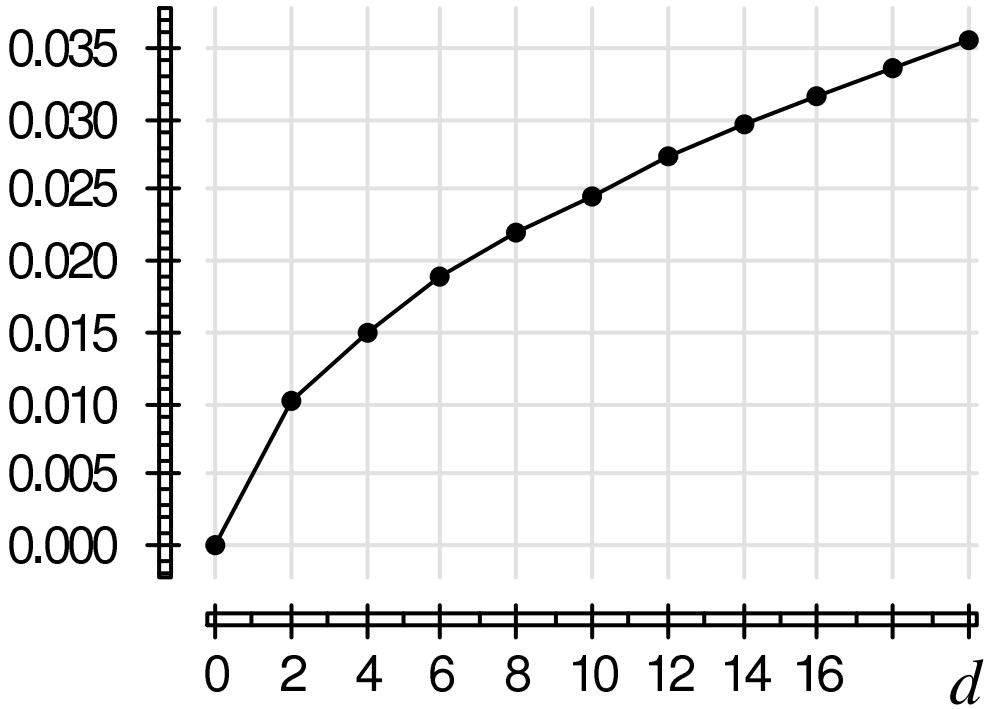


Рис. : Зависимость средней переобученности $\bar{\delta} = E_X \delta(\mu X, X)$ от d.

Сходство алгоритмов - пример 2

B_r^m — центральный слой шара (компактное множество),
R_n^m — алгоритмы с m ошибками; ошибки расположены случайно.

r	$\lvert B_r^m \rvert$	$\lvert R_n^m \rvert$	$\bar{\delta}$
2	401	2	0.079
4	35.501	7	0.160
6	1.221.101	39	0.240
8	20.413.001	378	0.319

Таблица : Сравнение $\lvert R_n^m \rvert$ и $\lvert B_r^m \rvert$ при $L = 50$, $\ell = 25$, $m = 10$

Задача: построить оценку, учитывающую сходство алгоритмов!

''Ингредиенты'' новой оценки:

1. разбиение множества алгоритмов на кластеры;
2. расширение кластера до слоя шара;
3. учёт расслоения;
4. эвристика: учёт размера каждого кластера.

Theorem (Фрей, Толстихин)

Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ *— разложение* A *на кластеры с равным числом ошибок. Пусть* $d_i = \sup\limits_{a,b \in A_i} \rho(a,b)$ *— хэммингов диаметр кластера* A_i*. Тогда*

$$Q_\varepsilon(A) \le \sum_{i=1}^{t} Q_\varepsilon(A_t) \le \sum_{i=1}^{t} Q_\varepsilon(B^m_{\lfloor d_i/2 \rfloor}), \text{ где}$$

$$Q_\varepsilon(B^m_r) = H^{\ell,m}_L(s_d(\varepsilon) + \lfloor r/2 \rfloor) \cdot [m \ge \varepsilon k].$$

''Ингредиенты'' новой оценки:

1. разбиение множества алгоритмов на кластеры;
2. расширение кластера до слоя шара;
3. учёт расслоения;
4. эвристика: учёт размера каждого кластера.

Theorem (Фрей, Толстихин)

Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ *— разложение* A *на кластеры с равным числом ошибок. Пусть* $d_i = \sup\limits_{a,b \in A_t} \rho(a,b)$ *— хэммингов диаметр кластера* A_t*. Тогда*

$$Q_\varepsilon(A) \leq \sum_{i=1}^{t} Q_\varepsilon(A_t) \leq \sum_{i=1}^{t} Q_\varepsilon(B^m_{\lfloor d_i/2 \rfloor}), \text{ где}$$

$$Q_\varepsilon(B^m_r) = H^{\ell,m}_L(s_d(\varepsilon) + \lfloor r/2 \rfloor) \cdot [m \geq \varepsilon k].$$

''Ингредиенты'' новой оценки:

1. разбиение множества алгоритмов на кластеры;
2. расширение кластера до слоя шара;
3. учёт расслоения;
4. эвристика: учёт размера каждого кластера.

Theorem (Фрей, Толстихин)

Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ *— разложение* A *на кластеры с равным числом ошибок. Пусть* $d_i = \sup\limits_{a,b \in A_i} \rho(a,b)$ *— хэммингов диаметр кластера* A_i*. Тогда*

$$Q_\varepsilon(A) \le \sum_{i=1}^{t} Q_\varepsilon(A_t) \le \sum_{i=1}^{t} Q_\varepsilon(B^m_{\lfloor d_i/2 \rfloor}), \text{ где}$$

$$Q_\varepsilon(B^m_r) = H^{\ell,m}_L(s_d(\varepsilon) + \lfloor r/2 \rfloor) \cdot [m \ge \varepsilon k].$$

''Ингредиенты'' новой оценки:

1. разбиение множества алгоритмов на кластеры;
2. расширение кластера до слоя шара;
3. учёт расслоения;
4. эвристика: учёт размера каждого кластера;

Theorem (Фрей, Толстихин)

Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ *— разложение* A *на кластеры с равным числом ошибок. Пусть* $d_i = \sup\limits_{a,b \in A_i} \rho(a,b)$ *— хэммингов диаметр кластера* A_i, q_i *— неполноценность кластера* A_i. *Тогда*

$$Q_\varepsilon(A) \le \sum_{i=1}^{t} \frac{C_{L-q_i}^{\ell}}{C_L^{\ell}} Q_\varepsilon(A_t) \le \sum_{i=1}^{t} \frac{C_{L-q_i}^{\ell}}{C_L^{\ell}} Q_\varepsilon(B^m_{\lfloor d_i/2 \rfloor}), \text{ где}$$

$$Q_\varepsilon(B_r^m) = H_L^{\ell,m}(s_d(\varepsilon) + \lfloor r/2 \rfloor) \cdot [m \ge \varepsilon k].$$

''Ингредиенты'' новой оценки:

1. разбиение множества алгоритмов на кластеры;
2. расширение кластера до слоя шара;
3. учёт расслоения;
4. эвристика: учёт размера каждого кластера;

Theorem (Толстихин, Фрей 2012)

Пусть B — множество алгоритмов с равным числом ошибок на полной выборке. Тогда среднюю ВП по всем подмножествам $A' \subset B$ фиксированной мощности $d = |A'|$ даётся выражением:

$$\bar{Q}_\varepsilon(B,d) = \frac{1}{C^d_{|B|}} \sum_{\substack{A' \subset B: \\ |A'|=d}} Q_\varepsilon(A') =$$

$$= 1 - \frac{1}{C^\ell_L} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \frac{C^d_{N_\varepsilon(B,X_\tau)}}{C^d_{|B|}}.$$

Точность новой оценки вероятности переобучения

1. Оценки Q и Q_s — без учёта сходства,
2. Оценки Q' и Q'_s — с учётом сходства, но без учета мощности кластеров A_i,
3. Оценки Q'' и Q''_s — с учётом сходства и мощности кластеров A_i.

			Без расслоения			С расслоением		
задача	метрика	Q_{cv}	Q	Q'	Q''	Q_s	Q'_s	Q''_s
echo-card	сред.переобуч.	0.033	0.246	0.204	0.185	0.216	**0.172**	**0.154**
	корреляция	1.000	0.619	0.601	0.500	0.813	**0.840**	**0.835**
hepatitis	сред.переобуч.	0.028	0.190	0.193	**0.165**	0.170	0.167	**0.149**
	корреляция	1.000	0.787	0.755	0.752	**0.792**	0.768	**0.818**
heart dis.	сред.переобуч.	0.013	0.132	0.107	**0.081**	0.124	0.108	**0.085**
	корреляция	1.000	0.716	**0.729**	0.681	0.722	**0.728**	0.714
labor	сред.переобуч.	0.066	0.405	0.347	0.333	0.367	**0.328**	**0.311**
	корреляция	1.000	0.644	0.622	0.636	**0.678**	**0.661**	0.659

Использование оценок для отбора конъюнкций

				Без расслоения			С расслоением		
задача	выборка	*Orig*	Q_{cv}	Q	Q'	Q''	Q_s	Q'_s	Q''_s
echo-card	обучение	0.1	0.1	0.7	0.5	0.3	0.2	0.2	0.3
	контроль	2.6	2.5	4.0	3.8	3.6	**1.0**	**1.0**	**1.0**
hepatitis	обучение	3.5	3.8	8.2	7.6	7.6	7.8	8.1	7.3
	контроль	19.2	18.5	**18.0**	19.0	19.2	**18.1**	18.5	18.7
heart dis.	обучение	8.2	11.1	10.8	11.2	10.6	10.6	10.8	10.1
	контроль	**18.7**	**18.7**	**18.7**	18.9	18.8	**18.7**	18.8	18.8
labor	обучение	0.5	0.9	1.7	1.8	1.8	1.5	1.5	1.0
	контроль	11.2	9.3	12.3	12.8	12.7	11.3	**10.8**	**10.6**

Таблица : Средняя частота ошибок (в процентах) на обучающей и тестовой выборке по различным задачам и различным методом контроля переобучения. Столбец *Orig* получен для не-модифицированного критерия информативности.

Сравнение с PAC-bayes оценками

Задача	Переобуч.	Монте-карло	Комб.оценка	PAC DI	PAC DD(*
faults	0.014	0.012	0.071	1.136	1.13
glass	0.069	0.111	0.093	1.085	0.72
Ionosphere	0.145	0.075	0.048	1.146	1.07
Liver dis.	0.022	0.079	0.221	1.151	1.08
Optdigits	0.008	0.005	0.235	1.083	0.69
pageblocks	0.003	0.005	0.107	0.436	0.19
pima	0.004	0.033	0.127	0.806	0.75
Sonar	0.328	0.133	0.134	1.343	1.34
statlog	0.006	0.012	0.251	1.126	0.87
waveform	0.005	0.005	0.213	0.412	0.35
Wdbc	0.060	0.045	0.014	1.199	0.82
Сред.завыш.	1.000	1.914	18.39	86.70	65.7

Таблица : Среднее уклонение частоты ошибок между контрольной и обучающей выборкой (10x кросс-валидация), и различные оценки вероятности переобучения для логистической регрессии.

(*) Dimensionality Dependent PAC-Bayes Margin Bound. Chi Jin, Liwei Wang. NIPS-2012.

Результаты выносимые на защиту:

- РМЭР и метод орбит для вывода оценок переобучения;
- общая оценка вероятности переобучения, учитывающая сходство алгоритмов семейства с помощью кластеризации;
- применение комбинаторных оценок вероятности переобучения в простом голосовании конъюнкций;
- экспериментальный модуль для сэмплирования семейств линейных классификаторов;
- экспериментальное сравнение комбинаторных оценок с оценками PAC Bayes (для логистической регрессии).

Не выносим на защиту:

- метод порождающих и запрещающих множеств для РМЭР;
- другие рандомизации (на метках целевых классов и независимые перестановки на векторах ошибок);
- учет верхней связности в комбинаторных оценках вероятности переобучения;